OPERA II — APPENDICI AVANZATE (Vol. 1)

Validazione causale, valutazione offline delle policy, controllo drift, envelope di sicurezza, evidenza auditabile.

Autore: Pierernesto Di Paolo

Anno: 2026

Tutti i diritti riservati

NOTE LEGALI (richiamo)

Identiche a Opera II: divieto di implementazione/uso clinico senza autorizzazione scritta dell'Autore.

Questo volume estende metodi e procedure: è parte del corpus proprietario dell'Autore.

INDICE

Appendice II-A — Politiche, controfattuali e regret osservabile (impostazione causale)

Appendice II-B — Off-Policy Evaluation: stime conservative del guadagno strutturale

Appendice II-C — Drift e “shutdown rules”: criteri formali per sospensione automatica

Appendice II-D — Safety envelope: vincoli di rischio e dominio operativo

Appendice II-E — Validazione clinica a livelli (V0→V3) e failure modes

APPENDICE II-A — CONTROFATTUALI E REGRET OSSERVABILE

II-A.1 Problema fondamentale

Il regret vero richiede confrontare $y(z,u^*(z))$ e $y(z,u_{pop})$ sullo stesso paziente, cosa non osservabile simultaneamente.

II-A.2 Formalismo minimo

Definisci esiti potenziali (potentials):

$Y(u)$ = esito se si applica decisione u.

Per il paziente i con stato z_i, osserviamo solo $Y(u_{obs})$.
Il regret richiede stime controfattuali $Y(u_{alt})$.

II-A.3 Conseguenza

Qualunque valutazione empirica del guadagno strutturale richiede una procedura controfattuale o un disegno sperimentale (randomizzazione / trial adattivo / stratificazione).

Questo non è un difetto della teoria; è limite epistemico inevitabile.

APPENDICE II-B — OFF-POLICY EVALUATION (OPE) CONSERVATIVO

II-B.1 Setup

Dati osservazionali $D = \{(W_i, u_i, y_i)\}$ con politica di logging $\pi_b(u|W)$.
Vogliamo stimare il valore di una nuova policy π_e (es. SSOC):

$V(\pi) = E[\ y \mid \text{policy}\ \pi\ ]$

II-B.2 IPW (inverse propensity weighting)

Stima base:

$V_hat_IPW(\pi_e) = (1/n)\ \Sigma_i\ [\ y_i * \pi_e(u_i|W_i)\ /\ \pi_b(u_i|W_i)\ ]$

Problemi: varianza alta, support mismatch.

II-B.3 Stima doubly robust (DR)

Con modello $y_hat(W,u)$:

$V_hat_DR(\pi_e) = (1/n)\ \Sigma_i\ [\ y_hat(W_i, \pi_e) + (\pi_e(u_i|W_i)/\pi_b(u_i|W_i)) * (y_i - y_hat(W_i,u_i))\ ]$

II-B.4 Forma “safety-first” (lower confidence bound)

Per uso serio (non retorico), si lavora con bound inferiore:

$LCB(\Delta V)$ = lower bound statistico su $V(\pi_e) - V(\pi_b)$

Se $LCB(\Delta V) \leq 0 \rightarrow$ non si autorizza passaggio policy.

Questo è coerente con filosofia bunker: non si “assume beneficio” senza margine.

APPENDICE II-C — DRIFT, SHUTDOWN, E RE-CALIBRAZIONE

II-C.1 Drift su Z e su outcome

Monitorare distribuzione di z_hat e residui (y − y_hat).

Definire statistiche:

D_Z(t) = distanza tra distribuzione corrente e baseline (es. MMD o Wasserstein)

D_y(t) = drift su residui

II-C.2 Regole di stop (deterministiche)

Se:

D_Z(t) > τ_Z oppure D_y(t) > τ_y oppure fallimenti solver > τ_s

allora:

SSOC entra in “SAFE MODE”:

u_final = u_pop

e registra evento in audit chain (Appendice I-bis).

APPENDICE II-D — SAFETY ENVELOPE (VINCOLI DI RISCHIO)

II-D.1 Vincolo di rischio tail

Richiedere:

$P_rare(z_hat, u_struct) \leq P_rare_max$

Se violato → fallback.

II-D.2 Vincolo di tossicità

Richiedere:

$T(z_hat, u_struct) \leq T_max$

II-D.3 Dominio operativo

SSOC opera solo se:

- qualità input sufficiente,
- stima Σ_e affidabile,
- δ (errore modello) sotto soglia,
- drift sotto soglia.

Questo impedisce uso fuori dominio.

APPENDICE II-E — VALIDAZIONE A LIVELLI E FAILURE MODES

II-E.1 Livelli

V0: retrospettivo + OPE conservative (solo raccomandazioni offline)

V1: prospettico osservazionale (raccomandazione non esecutiva)

V2: trial DSS controllato

V3: eventuale percorso regolatorio (se e dove applicabile)

II-E.2 Failure modes principali

- mis-specificazione di y_hat
- stima tonica inaffidabile (Σ_e sottostimata)
- shift di popolazione
- leakage / bias di selezione
- support mismatch per OPE
- instabilità numerica

II-E.3 Mitigazioni

- fallback automatico
- audit immutabile
- stop rules
- monitor drift
- LCB per decisioni di policy change
- revisione umana obbligatoria